# A Platform for Managing Term Dictionaries for Utilizing Distributed Interview Archives

Kenro Aihara and Atsuhiro Takasu

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{kenro.aihara,takasu}@nii.ac.jp

**Abstract.** This paper proposes a platform that aims to support the whole process and facilitate archiving tasks at museums and galleries.
When we try to preserve tacit knowledge or skills of artists or masters by interview and utilize interview videos, it is important to support the whole process of archiving; capturing interview, digitizing it to a machine readable format, authorizing metadata, and exporting them to be accessed from others.
To interoperate distributed archives, rich meta information is required. In this paper, we focus our tools to facilitate authoring metadata of interview, such as transcription and annotation task for interview, with speech processing or natural language processing functions. The important point now is that the performance of the speech processing or text processing for specific domain depends on well-fitted language models to it. We, therefore, design our tools as an networked software package that can download up-to-date term dictionary and language models for speech processing or text processing from a server of the platform. Users can utilize tools with updated models. Meanwhile, created metadata including terms occurred in interview can be uploaded to the server and they will be used for update of the models hereafter.
In this paper, we first overview our project. Then, we propose an platform and our tools to be distributed.

## 1 Introduction

For the field of arts and crafts, domain knowledge or individual knowledge, such as tacit knowledge or skills of artists or masters, are not inherited for lack of successors or effectual methods of communications to the next generation. We, therefore, must consider how to support patrimony of such knowledge or skill of human creativity.

We suppose that text is not enough because skill and knowledge are sometimes beyond description and we need more excellent form for recording them. Therefore interviews with masters and artists are effective way to record their skill and knowledge instead of textbook type description. We think that interview videos can keep not only master's words but also his/her gestures, atmosphere, savvies, and context of them. We have been constructing an interview video

archive for preserving such knowledge of masters in the field of lacquer arts and crafts for this purpose.

We must consider to integrate such interview archives efficiently when archives are created individually and its number gets grown. This paper focuses this issue and proposes our approach for improve information integration by sharing fundamental resources.

In this paper, we propose a platform for management of term dictionary for utilizing distributed interview archives. Section 2 describes the background of this research. In Section 3 , our project called MONO is overviewed. Then, Section 4 describes our platform for managing term dictionary. Conclusions are given in Section 5.

## 2 Background

### 2.1 Sharing Skill and Knowledge of Artistic Creativity

Sharing skill and knowledge of master workmen and artists is one of important and challenging issues. Usually disciples acquire the skill and knowledge by oral communication with masters and watching the master's works. They, therefore, can be conveyed to limited number of disciples and they are sometimes lost when masters and artists pass away. A networked mechanism for sharing skill and knowledge must be needed to preserve them and convey it to a large amount of people.

Since skill and knowledge inherent in masters and artists, the first step to construct such a mechanism is to externalize them and represent them in an appropriate form. We want to emphasize that this means that this kind of knowledge sharing needs not only an efficient file sharing mechanism but also support of digitization to externalize them.

Text is not appropriate because skill and knowledge are sometimes beyond description and we need more excellent form for recording them. Interviews with masters and artists are effective way to record their skill and knowledge instead. It can record various kinds of information such as emotional behavior, procedure of creative activity as well as textual information in the form of conversation. Interview has another advantage that it enables to obtain the information from masters and artists quickly without heavy mental load. It will take a year or more for masters to write a book of their skill and knowledge. However it takes only several hours to have interviews. The latter advantage of interview is very effective to solve the bottle-neck problem of information capturing not only for the skill and knowledge of masters and artists but also for externalizing the knowledge of human beings in many fields.

### 2.2 Language Models for Processing

When we try to integrate distributed resources, we must consider two aspects: data type and domain.

In recent competitive solutions for multimedia retrieval, textual metadata or annotation is attached first to each non-textual data and then an index for image retrieval is made[3]. This suggests that creating appropriate textual metadata corresponding multimedia data is critical. Once textual metadata is attached, multimedia contents can be handled as the same as text.

In text retrieval, feature of data is usually defined with frequency of term occurrence, represented by Salton's TF-IDF weighting[5]. We must consider that controlled term sets are necessary if we try to improve efficiency of such weighting. In other words, sharing domain specific terms and its maintenance is critical to integrate distributedly generated contents. Construction of domain ontology can be regarded as one of approaches for this purpose. However, we cannot utilize such effective and flexible ontologies for now.

On the other hand, recent researches on both natural language processing (NLP) and speech recognition get advanced in corpus-based approach, which applies language models learned from data collection.

It is important to note that efficient processing for text and multimedia data depends on shared terms and language models corresponding each domain. Especially, they are necessary if the domain is highly technical and there are relatively small amount of digital contents.

We, therefore, have been developing a platform for sharing fundamental resources of specific domain, such as technical terms.

## 3 Overview of MONO project

In this section, we describe our project called MONO[1] to archive skill and knowledge of artists [1]. In this project, we adopted interview videos as the main form of information representation. To the best knowledge of authors, only a few digital libraries handle interview videos such as a VHF's archive of interviews to survivors of the Holocaust [4]. The MONO project will clarify new technical challenges and derive a new framework for constructing archives.

### 3.1 Purpose

The purpose of MONO project is to create a platform for archiving skill and knowledge of artists in the field of the arts and crafts and providing it for students in this field as well as visitors to museum. The goals of this project are:

- Constructing an archive of skill and knowledge of masters and artists to preserve the knowledge otherwise it may be lost with masters in near future.
- Developing information processing technologies for archiving, e.g., editing interview videos and attaching metadata to them.
- Developing an interview archive utilization methods such as efficient interview video search and effective browsing methods.

[1] General meaning of MONO is an object in Japanese, but this word also means an valuable object created by an excellent workman or artist.

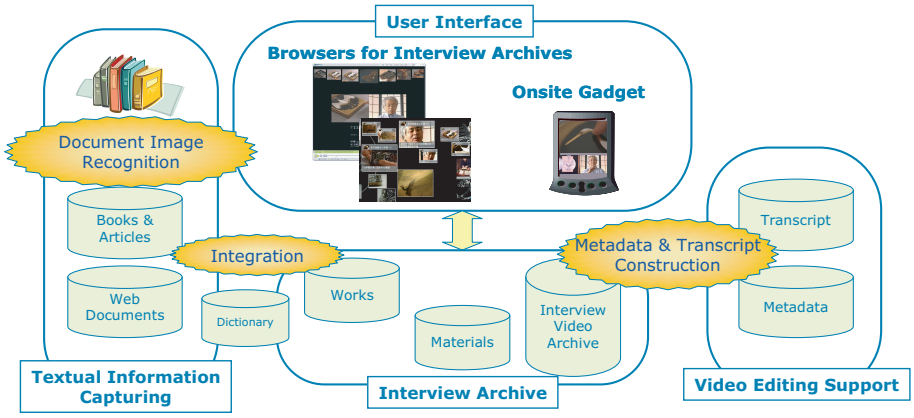


**Fig. 1.** An Overview of MONO project

– Constructing a test bed for multimedia processing technology such as speech recognition, video processing and information extraction.

For the test bed construction, we are attaching metadata such as transcripts for interview videos, scene change of videos, topic change of the interviews for evaluation.

### 3.2 Architecture

A system of the MONO project consists of data collection and three modules as shown in Fig. 1.

**Collection** The system contains various kinds of information. They are categorized into two groups: interview videos and complementary textual information. Interview videos is the main part of the collection. In the interview, masters and artists explain their creative works using their artifacts. Currently the collection contains 21 interviews of artists and masters in the field of lacquer arts and crafts. In this project, we have been recording some interviews with two cameras. One is for recording interviewee's face and the other is for his/her hands or works which he/she is handling. The format of interview video is AVI with the resolution of 720 $\times$ 480 frame, 29.97 frames per second. Total length of interviews is 2,088 minutes. We continue interviews and plan to collect 30 interviews within this year.

Although an interview is a very effective way to record the skill and knowledge, it is sometimes too specific for users who do not have sufficient knowledge in this field. In order to help users to comprehend this field and understand the interview contents well, the collection contains complementary textual information such as dictionary, articles, books and catalogs of exhibitions.

**Fig. 2.** Snapshot Images of "Lacquer Arts and Crafts" Collection

**Functional Modules** Current system has three functional modules: a video editing support module (VESM), a textual information capturing module (TICM) and a user interface module (UIM). VESM and TICM provide archiving functions and UIM provides an information utilization function.

Usually the interview contains several topics. In editing process, interview videos are segmented into sections and each section is further segmented into scenes. On the other hand, conversation in the interview video is converted to transcript. Video segmentation and transcription generation are basically done by hand. This part is most labor intensive in archive construction. Cost reduction of this process is a key to construct large archive. Currently VESM supports this process in two ways: scene change detection and transcript generation. There are several techniques for detecting scene change in videos. These techniques usually use the change of color frequency distribution between frames. However, in interview videos, the color frequency distribution does not always change drastically even if the scene changes. VESM detects and provides the candidate scene changes for a video editor based on the speech signals. Second, VESM has speech recognition ability. It recognizes the conversation in interview videos. Since the recognition accuracy is not high enough to use the resultant transcript as the final one, editors need to correct the recognition error. As for the speech recognition, the goal of this project is to establish a video search technique without manually produced transcript.

As described above, our collection contains complementary textual information as well as interview videos. TICM supports to capture these textual information. Currently TICM has the following functions. Books and articles sometimes need digitization from printed paper. TICM has OCR and document image analysis submodule [7] to digitize them. In order to support dictionary

edition, TICM has a natural language processing module to extract important words from documents based on a statistical measure [2]. Dictionary editors can select entry words from these candidates in dictionary edition. We are now developing a collaborative dictionary editing environment where dictionary editors collaboratively write the meaning of the entry words and set links to related words.

UIM provides effective access to the archive.

**Technical Challenges** Interview video archive is a kind of video archive and it has same technical problems as general video archives. In this subsection, we focus on the technical problems specific to interview videos.

*Metadata Construction* Metadata is very important to utilize videos. Our system has two level structure of meta data for interview video: *scenes* and *sections*. A section is a logically minimum unit of video corresponding to one verbal phrase. And a scene is a continuous parts of video. This metadata structure is similar to general video archives. However, in the case of interview video, we can control the interview in the section level in some extent. Currently we are designing a standard interview manual. It will support to make section level metadata. In VHF's interview archive [4], pre-interview questionary is carried out. Resultant documents are attached to interviews and used in interview retrieval. In our system, the manual is combined to interview video tightly and it will be used to retrieve interview videos at the section level. The point of metadata construction for interview video is pre-interview story configuration which derives homogeneous video archive.

As mentioned above, in the interview video, it is difficult to extract scenes based on color distribution of frames because the main object of the interview is a master or an artist, and this feature tends not to change during shooting. Currently our system detects the scene changes based on the speech signals. However, in order to reduce the editing cost we need a more accurate method. In order to improve the accuracy of scene change detection, we have been developing an authoring tool to cut scenes detailed later and also are planning to develop an equipment with which interviewer can record marks at scene change during the interview.

*Speech Processing and Recognition* In the interview videos, the conversation is the dominant information sources. Therefore, speech processing technology is much more important than video processing technology. Accuracy of current speech recognition is not sufficient enough, and recognized textual information is too erroneous. Therefore we need a robust technique to utilize conversation in the form of speech signals.

*Information Integration* In interview archives, complementary textual information is especially important to provide comprehensive view of the archive for users because the information in the interview is usually very specific. Therefore,

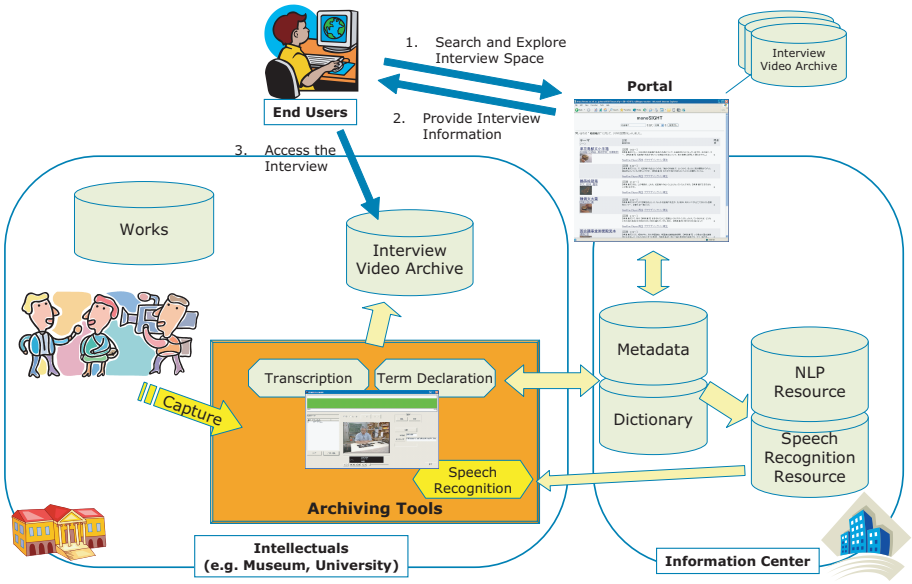


**Fig. 3.** Overview of the Proposed Platform

related textual information such as articles, books, exhibition catalog and web pages should be associated with interview video. From the technical viewpoint, transmedia information integration technology is a key to solve this problem. In MONO project, we first construct a dictionary, then associate related textual information with scenes and sections of interview videos using the dictionary as a base of information integration.

As mentioned in Section 3.2, this paper focuses a collaborative dictionary edition in the MONO project.

## 4 Proposed System

### 4.1 Overview

Our methodology aims not only at construction of portal site but also at supporting capture of digital contents transformed from interview videos with intellectuals. We assume that a reciprocal relation between portal and local repositories must be needed for successful portal construction.

Fig. 3 illustrates an overview of our proposed platform for construction of portal service and local repositories.

We provide a support tool package for archives for local repository sites, such as museum. The package includes manuals for interview, software to edit videos and to create metadata that includes transcripts of interview, and server program for online service.

The orange box in Fig. 3 indicates the software to support transcription and editing term dictionary. While terms in transcripts can be tagged manually in metadata creation phase by users, such as curator at museum, terms and description about them also should be added into local term dictionary.

In addition, the package also includes speech recognition module. Unfortunately performance of current speech recognition tools depends on the quality of language models and dictionaries. Our portal of "Information Center" in the figure,therefore, provides language models for speech recognition which are incrementally constructed with collected metadata and term dictionaries from local repositories. Users can download up-to-date data for speech recognition and apply it to their videos when they transcribe them.

The portal site can aggregate high quality metadata and language resource from each local repository. Aggregated language resources can be used not only for speech recognition but also natural language processing or information retrieval.

At first, a user has to register him/herself at the portal site and then download the archiving package.

After creation of an archive, the local site can provide its own video service. Typically, videos and metadata are integrated as SMIL contents. SMIL enables us to describe the temporal behavior of a multimedia presentation, associate hyperlinks with media objects and describe the layout of the presentation on a screen.[2] Fig. 4 illustrates a snapshot image of browsing interview video as SMIL content. The window of the SMIL browser contains multimedia objects as follows:

- the scene being played in the middle of the window
- caption under the current scene, which scrolls up synchronously
- sequence of thumbnails of scenes in chronological order on the top, while the current scene is located in the middle of the sequence

End users can be navigated to an appropriate archive when they search for contents at the portal.

We think that the sequence of scenes can help you grasp the context of the interview intuitively. It is obvious that scrolling caption is helpful to understand more correctively and easily than without it. Words in interviews often includes technical terms.

### 4.2 Archiving

**Metadata Structure** For interview videos, we define the whole interview as *commentary*. One commentary can includes some comments about one or more works or topics. One commentary consists of some *themes*. Theme is a time region of the interview and usually corresponds to one work or topic. We, therefore,

[2] SMIL, the Synchronized Multimedia Integration Language, is a standard language for simple authoring of interactive audiovisual presentations. W3C published SMIL 2.0 Recommendation in August 2001. http://www.w3.org/TR/smil20/


**Fig. 4.** Accessing Interview Video with an SMIL browser

regard a theme as a comment of each work. One theme consists of some *scenes* which are continuous parts of recorded video. And also one theme is segmented into *sections* by verbal gap or topical change. Section, therefore, is an atomic unit of contents here. Fig. 5 illustrates that one theme consists of one or more scenes of multiple cameras.

Utterance of interviews is transcribed and used as caption. This transcribed text is also used for indexing each section.

**Archiving Interview** Fig. 6 shows process flow of archiving.

At first, a user of this annotation tool captures and digitize interview videos, and the tool shows possible points of scene change. The user has to fix the points interactively.

Next, the user needs to segment each scene into sections. Viewing candidates of section according to verbal gap, the user fixes sections.

Then, the tool runs speech recognition function with language models provided at the portal to show a candidate of transcript for each section. While the user edits transcripts, he/she also declares terms by tagging and its description. The tool highlights terms which are already registered in the current dictionary to facilitate the term declaration task.

After the authoring process, the tools produces an interview video archive including videos and metadata and also uploads metadata with local term dic-

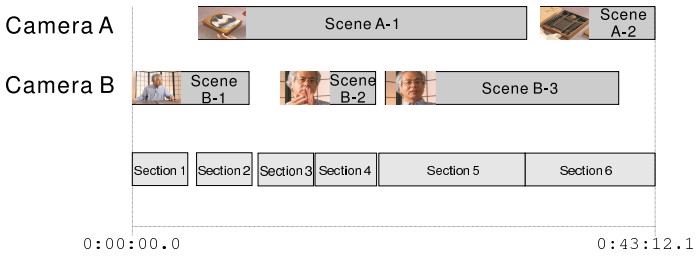


**Fig. 5.** Theme

tionary to the server of the portal. The gray parts in Fig. 6 indicates shared resources.

### 4.3 Shared Language Models

In order to support dictionary edition, we first have to create an initial dictionary. Books and articles sometimes need digitization from printed paper. TICM of Fig. 1 has OCR and document image analysis submodule [7] to digitize them. TICM has a natural language processing module to extract important words from documents based on a statistical measure [2]. Dictionary editors can select entry words from initial candidates in dictionary edition.

The models can be updated when local term dictionaries are uploaded.

Metadata including transcripts and term dictionary can be used as a fundamental resource of its own field; lacquer arts and crafts. It may be exploited in natural language processing of search engines of the portal.

## 5 Conclusion

This paper overviews our project called MONO which aims to reveal a mechanism for sharing knowledge or skill of artists or masters. In particular, we focus maintenance of shared term dictionary of the proposed platform in this paper because sharing fundamental resources for natural language processing and speech recognition is critical for integration of distributed contents.

Distribution of prototype tools and feasibility studies are future issues.

## Acknowledgments

This study is partly supported by Grand-in-Aid Scientific Research on Priority Area “Informatics” (Area #006).

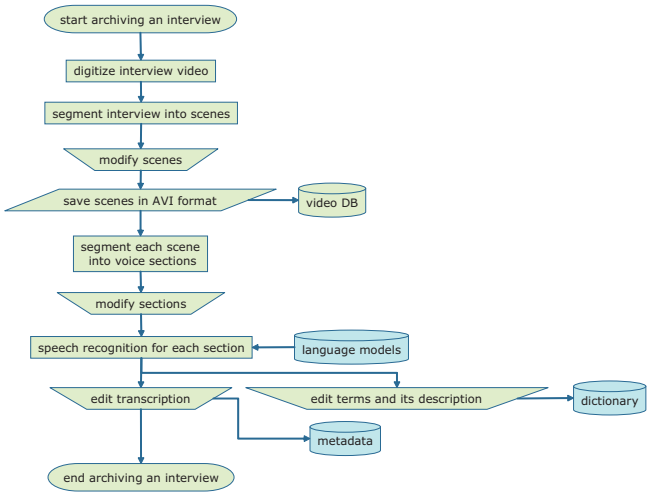


**Fig. 6.** Flowchart of Archiving

## References


1. K. Aihara and A. Takasu. A reciprocal platform for archiving interview video about arts and crafts. In *Joint Conference on Digital Libraries*, page 363, June 2005.
2. A. Aizawa. An information-theoretic perspective of TF-IDF measures. *Information Processing and Management*, 39(1):45–65, 2003.
3. P. Clough, M. Sanderson, and H. Müller. The CLEF cross language image retrieval track (ImageCLEF) 2004. In *Working Notes for the CLEF 2004 Workshop*, 2004.
4. S. Gustman, S. Soergel, D. W. Oard, and W. J. Byrne. Supporting access to large digital oral history archives. In *Joint Conference on Digital Libraries*, pages 18–27, 2002.
5. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1998.
6. S. Srinivasan and D. Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–87, 2000.
7. A. Takasu. Probabilistic interpage analysis for article extraction from document retrieval. In *Proceedings of the 14th International Conference on Pattern Recognition*, pages 932–935, 1998.
8. A. Takasu. Bibliographic attribute extraction from erroneous references based on a statistical model. In *Joint Conference on Digital Libraries*, pages 46–60, 2003.
9. A. Takasu and K. Aihara. DVHMM: Variable length text recognition error model. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume III, pages 110–114, 2002.
10. M. Wechsler, E. Munteanu, and P. Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–27, 1998.